We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,300
Open access books available

130,000
International authors and editors

155M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter

Introductory Chapter: Homology Modeling

Rafael Trindade Maia, Magnólia de Araújo Campos and Rômulo Maciel de Moraes Filho

1. Introduction

Proteins are macromolecules present in all living beings and perform a huge variety of complex and diverse functions and structures. They are polymers of amino acids synthesized in the cell of living organisms, also called polypeptides. Determining the three-dimensional structure of a protein is crucial for understanding its function. However, experimental techniques for structural elucidation such as X-ray critalography and nuclear magnetic resonance (NMR) are complicated and expensive [1]. In this context, computational techniques for building structural models are a very useful and viable alternative for different situations. Among computational techniques, homology modeling, also known as comparative modeling, is the most used *in silico* tool for obtaining structural protein models, achieving excellent results [2].

Proteins are organized at different levels of structural complexity: 1) primary structure; 2) secondary structure; 3) tertiary structure; 4) quarternary structure (**Figure 1**). The primary structure of a protein comprises the linear sequence of the amino acids that compose it, with one end containing the carboxyl group of the first amino acid in the chain (C-terminal) and with one end containing the amino group of the last amino acid in the chain (N -terminal). The primary structure of a protein can be represented by a pattern of letters that represents its peptide constitution (amino acids). The secondary structure of a protein is determined by the primary sequence, which is decisive in the arrangement of the monomers (aminoacids) with each other and with the solvent, forming standard structures in three groups: the turns, the helix and the β-leaves. The way in which these secondary structures are organized three-dimensionally in space is what is called a tertiary structure, which is associated with the biological function of the molecule in question. In multimeric protein complexes (dimers, trimers, tetramers, etc.) there is also the formation of the quarternary structure, which is the oligomeric state formed by the aggregation of these macromolecular compounds of tertiary structure.

There are three types of computational modeling for predicting protein structures: by *ab initio*/*De novo*, by *Threading* and by homology modeling. Homology modeling is based on the premise that the three-dimensional structure of a protein tends to be much more conserved than its primary structure. Therefore, changes in the sequence do not always change the structural domains of a protein, thus maintaining its original function. It is assumed that proteins from the same functional family maintain their structural domains, which allows the so-called comparative modeling (by homology). If two proteins are homologous, it means that they belong to the same genetic and functional family, and hypothetically, they have the same structural motifs. In the case of a specific protein that does not have an elucidated three-dimensional structure, but it is homologous to a protein with a

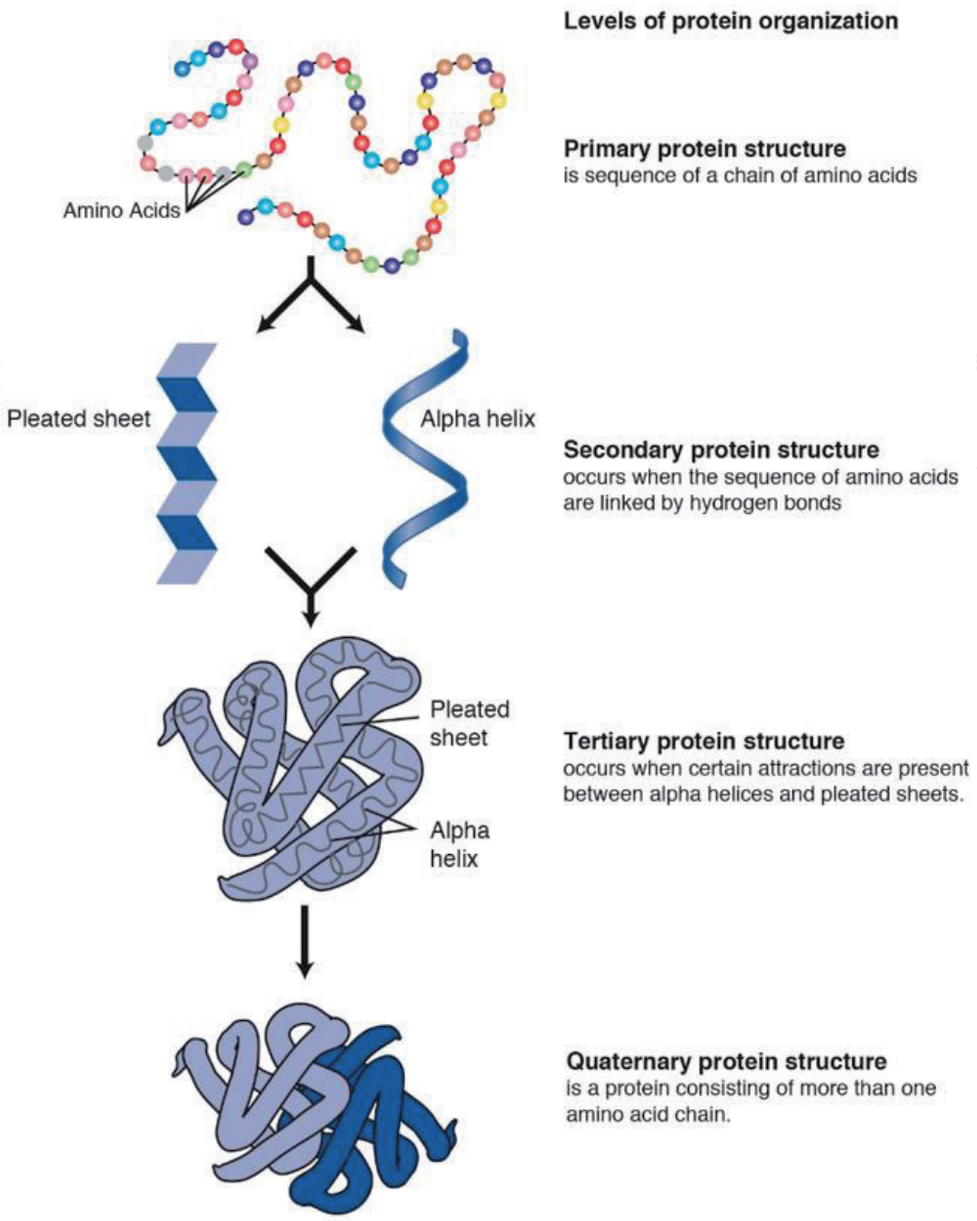

Figure 1.
Illustrative scheme for the structural complexity levels of proteins. Source: Google images.

solved structure, a three-dimensional model for the sequence can be built using the known structure as a template. As a rule, a minimum identity of 25% between the amino acids of two proteins is sufficient for the construction of models by homology. Sequence identities above generally 40% provides good models, while those above 50% tend to provide excellent theoretical structures [3].

However, in addition to the identity and similarity between the amino acids, other parameters must be observed when choosing a good template, such as the resolution in angstroms of the crystallographic structure and the percentage of alignment coverage (**Figure 2**). The lower the resolution of a structure, the better its quality. The average resolution of the structures available in the PDB (Protein Data Bank) is around 3.5 Ä, while structures below 2.0 Ä are considered to have excellent resolution and represent less than 10% of the entries in the PDB. The higher the percentage of coverage of the alignment between a target protein (protein to be modeled) and the template (mold), the better [4]. Coverage alignments above 90% of the residues tends to have high scores and are considered to be excellent (**Figure 2**).

Something important to note in alignments is the presence of sequence gaps. A gap between sequences means the absence of residues, that is, amino acids that

Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Yeast V-ATPase state 1 [Saccharomyces cerevisi…	baker's …	111	111	95%	8e-30	30.48%	233	3J9T_G
Cryo-EM structure of V-ATPase from bovine brain…	cattle	109	109	97%	5e-29	32.09%	226	6XBW_I
Mammalian V-ATPase from rat brain - composite …	Norway …	107	107	97%	2e-28	31.63%	226	6VQ6_I
Structure of E1-69 of Yeast V-ATPase [Saccharo…	baker's …	37.7	37.7	24%	0.001	37.74%	69	2KZ9_A

Figure 2.
Example of BLASTp alignment between a Leishmania infantum *ATP-synthase sequence against the PDB database. Values of the coverage percentage (red) and identity (black) of each alignment are highlighted. Source: Authors data.*

Score	Expect	Method	Identities	Positives	Gaps
43.1 bits(100)	6e-05	Compositional matrix adjust.	44/158(28%)	84/158(53%)	8/158(5%)

```
Query  29   QSSAFFGSMGCASALIFANLGSAYGTAKSGVGVAHLGILHAERIMRGIVPVVMAGILGIY  88
            S   + ++G A  +  + +G+A+G   +G  +   G+       + ++ ++   ++ IY
Sbjct  54   TSPYMWANLGIALCVGLSVVGAAWGIFITGSSMIGAGVRAPRITTKNLISIIFCEVVAIY  113

Query  89   GLIVSVIINNNIH--ADDNSYSA---FAGYLHFGAGLAAGLSSLAAGLSIGIAGDASVRA  143
            GLI++++ ++ +     +N YS     + GY  F AG+  G S+L  G+++G+  +  +
Sbjct  114  GLIIAIVFSSKLTVATAENMYSKSNLYTGYSLFWAGITVGASNLICGIAVGITGATAAIS  173

Query  144  YGKQEKIFVAMILMLIFAEALGLYGLIIALLMNNTAGK  181
                 +FV ++++ IF   LGL GLI+ LLM   AGK
Sbjct  174  DAADSALFVKILVIEIFGSILGLLGLIVGLLM---AGK  208
```

Figure 3.
Alignment between two proteins (query/Sbjct) showing the presence of 8 gaps (red) in three different sections (green). Source: Authors data.

have been deleted from some part of the sequence (**Figure 3**). The amount and size of gaps in an alignment is crucial to the final quality of the models. The greater the quantity and size of the gaps, the less reliable the models are and the greater is the chance of generating structural artifacts. Therefore, when choosing a template, it is essential that the researcher be aware about gaps presence in the sequences.

Once the template has been defined, we proceed to the stage of building the three-dimensional model. From specific programs and servers, the necessary files for modeling are submitted, which consists of the superimposition of the structural carbons of the target protein on the template protein, based on the alignment information to superimpose the equivalent amino acids. There are currently numerous free tools for building three-dimensional models (**Table 1**).

Nome	Tipo	Site
Modeler	Software	https://salilab.org/modeler/
Swiss-Model	Server	https://swissmodel.expasy.org/
Phyre2	Server	http://www.sbg.bio.ic.ac.uk/phyre2
Galaxy	Server	http://galaxy.seoklab.org/
RaptorX	Software/Server	http://raptorx.uchicago.edu/
CONFOLD	Software	https://github.com/multicom-toolbox/CONFOLD
ROBBETTA	Server	http://robetta.bakerlab.org/

Source: Google search.

Table 1.
Examples of free tools for building homology models.

2. Validation and refinement

Homology models are theoretical-computational approximations of the real protein structures, and therefore require validation and sometimes refinement and optimization. A very popular validation tool is the Ramachandran plot (**Figure 4**), which analyzes the stereochemical quality of protein structures.

The Ramachandran graph analyzes the conformations of the *phi* and *psi* angles of the peptide bonds, placing them in regions. Residues outside the permitted regions (outliers) are those that are in unfavorable configurations due to the collision between the atoms (steric shock). It preconizes that a good model should have at least 90% of its waste in favorable and permitted regions [5].

Other validation tools are energy assessments, both local and global ones. A tool for global assessment of the quality of a model is the server PROSA-web - Protein Structure Analysis (https://prosa.services.came.sbg.ac.at/prosa.php) [6, 7], which compares the energy of a structure with a database of proteins of equivalent size, solved experimentally, through the Z-score (**Figure 5**).

For local quality analysis, the application of the VERIFY3D server (https://servicesn.mbi.ucla.edu/Verify3D) is very useful. In this type of analysis it is possible to check the local quality, that is, for each residue of the model (**Figure 6**). With this, it is possible to identify specific regions of low quality for further adjustments.

For the models refinement, two techniques are particularly interesting: energy minimization and classical (atomistic) molecular dynamics. Energy minimization, also called optimization of geometry, aims to find a set of atomic coordinates of the structure that avoid bad contacts and reduce the potential energy of the system. There are some free servers available for energy minimization application in theoretical models, like YASARA [8] (http://www.yasara.org/minimizationserver.htm) and CHIRON [9] (https://dokhlab.med.psu.edu/chiron/). Molecular dynamics are extremely efficient for validating and refining theoretical models. This technique is based on the principles of Classical Mechanics and describes the atomic movements

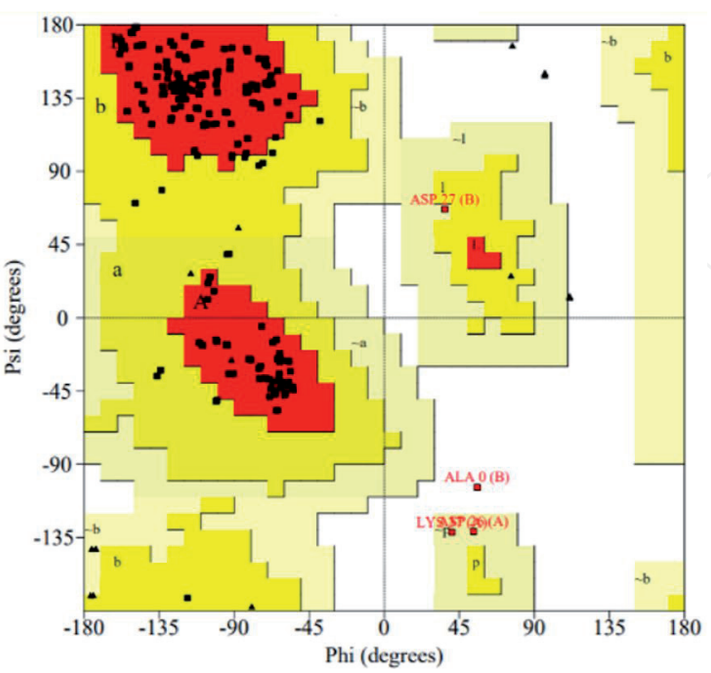

Figure 4.
Ramachandran graph for SARS-CoV-2 NSP9 replicase (PDB ID: 6w4b). In red, more favorable regions. In yellow and beige, regions allowed. In white, forbidden regions. Source: Authors data.

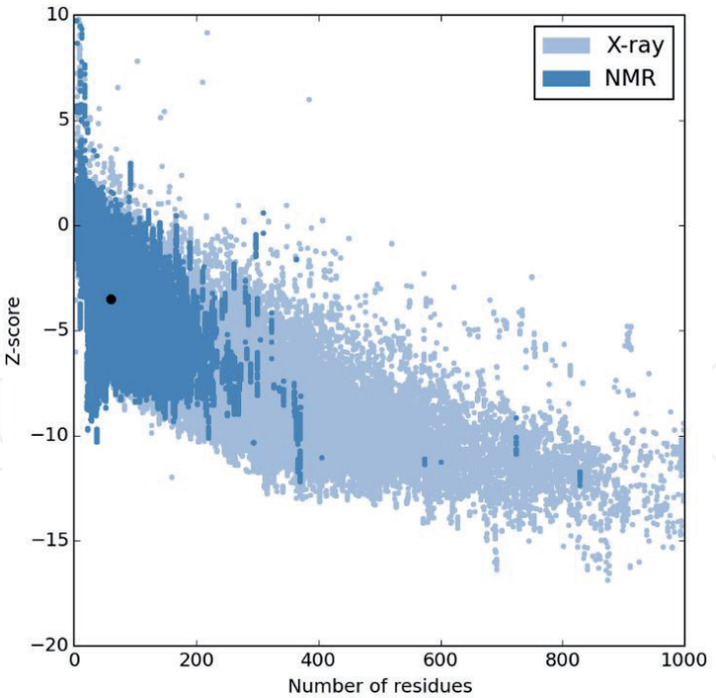

Figure 5.
Comparative graph of the Z-score energy. The black dot represents the position of the analyzed protein compared to equivalent size structures obtained by x-ray crystallography (light blue) and nuclear magnetic resonance (dark blue). Source: Authors data.

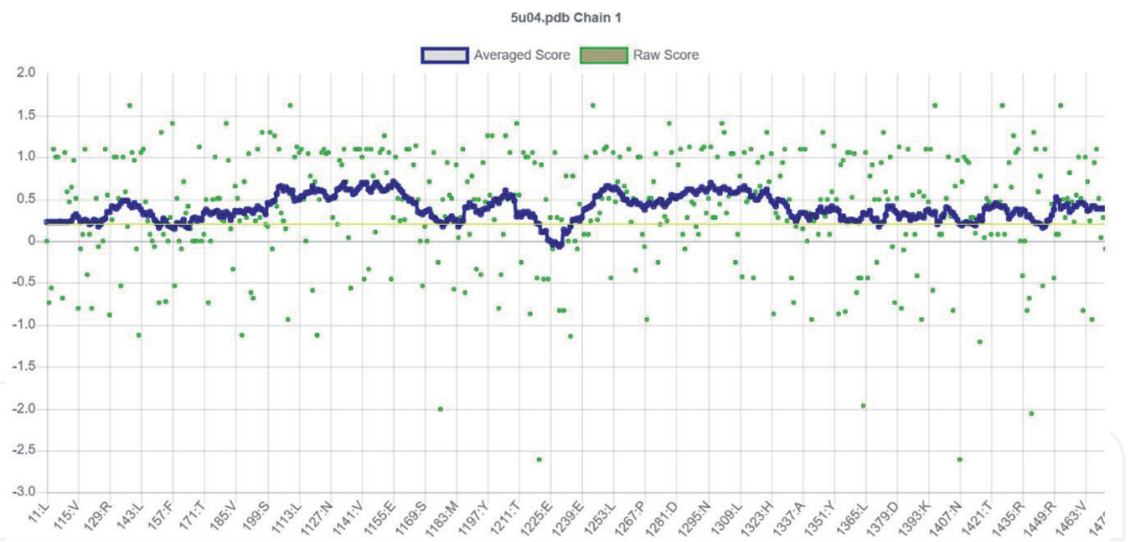

Figure 6.
Local ERRAT quality graph of a stretch from the NS5 enzyme from Zika virus. In blue, the average scores, in green, the raw scores. 93.93% of the residues have averaged 3D-1D score > = 0.2 (80% indicates good structures). Source: Authors data.

of a system through the integration of Newtonian equations of motion. Thus, a molecular dynamics simulation of 5–10 nanoseconds is one of the most effective techniques for optimization and validation of models by homology. For performing molecular dynamics calculations, software such as GROMACS [10] and NAMD [11] are useful. Once optimized and validated, the theoretical model can be used for several purposes, and can also be made available in public repositories, such as the PMDB - Protein Model DataBase (http://srv00.recas.ba.infn.it/PMDB/) and the SWISS-MODEL repository (https://swissmodel.expasy.org/repository).

3. Conclusions

Theoretical-computational models are fast, inexpensive and extremely versatile. There are countless possibilities for studies and uses of models by homology. These structures can be used for drug screening, docking studies, development of new drugs and vaccines, elucidation of binding sites (catalytic and allosteric), molecular dynamics simulations, quantum studies, biomolecule engineering etc.

The future of molecular modeling is fascinating and promising. With the advancement of computational tools, theoretical models tend to be increasingly accurate and reliable, contributing more and more to biological and biotechnological researches, in addition to integrating various areas of knowledge with bioinformatics and computational biology.

Acknowledgements

The authors are grateful to the Federal University of Campina Grande and to Federal Rural University of Pernambuco.

Conflict of interest

The authors declare no conflict of interest.

Author details

Rafael Trindade Maia[1*], Magnólia de Araújo Campos[2] and Rômulo Maciel de Moraes Filho[3]

1 Federal University of Campina Grande, Centro de Desenvolvimento Sustentável do Semiárido, Sumé, Paraíba State, Brazil

2 Federal University of Campina Grande, Centro de Educação e Saúde, Cuité, Paraíba State, Brazil

3 Federal Rural University of Pernambuco, Departamento de Agronomia, Recife, Pernambuco State, Brazil

*Address all correspondence to: rafael.rafatrin@gmail.com

References

[1] Luzzolino L, McGabe P, Prince S L, Brandenburg J G. Crystal structure prediction of flexible pharmaceutical-like molecules: density functional tight-binding as an intermediate optimisation method and for free energy estimation. Faraday Discuss., 2018, 211, 275.

[2] *Haddad Y*, *Adam V, Heger Z*. Ten quick tips for homology modeling of high-resolution protein 3D structures. *Plos Computational Biology*. *2020* 16(4): e1007449. doi:10.1371/journal.pcbi.1007449.

[3] Verli H. Bioinformática: da biologia à flexibilidade molecular. São Paulo, 2014. Ed. Sociedade Brasileira de Bioquímica e Biologia Molecular – SBBq.

[4] Leach AR. Molecular Modelling: Principles and Applications. Prentice Hall, 2001.

[5] Laskowski R A, MacArthur M W, Moss D S, Thornton J M (1993). PROCHECK - a program to check the stereochemical quality of protein structures. J. App. Cryst., 26, 283-291.

[6] Wiederstein, M. & Sippl, M.J. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research 35, W407-W410.

[7] Sippl, M.J. (1993). Recognition of Errors in Three-Dimensional Structures of Proteins.

[8] Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. Proteins. 2009;77 Suppl 9:114-22.

[9] Chiron : Ramachandran, S., Kota, P., Ding, F. and Dokholyan, N. V., PROTEINS: Structure, Function and Bioinformatics, 79: 261-270 (2011).

[10] Abraham, et al. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1-2 19-25.

[11] James C. Phillips, David J. Hardy, Julio D. C. Maia, John E. Stone, Joao V. Ribeiro, Rafael *C. bernardi*, Ronak Buch, Giacomo Fiorin, Jerome Henin, Wei Jiang, Ryan McGreevy, Marcelo C. R. Melo, Brian K. Radak, Robert D. Skeel, Abhishek Singharoy, Yi Wang, Benoit Roux, Aleksei Aksimentiev, Zaida Luthey-Schulten, Laxmikant V. Kale, Klaus Schulten, Christophe Chipot, and Emad Tajkhorshid. Scalable molecular dynamics on CPU and GPU architectures with NAMD. Journal of Chemical Physics, 153:044130, 2020. doi:10.1063/5.0014475